STA496 Forecasting Assignment

Talia Fabregas

August 2025

1 Assignment overview

Former President Joe Biden made the unprecedented decision to end his re-election campaign on July 21, 2024 following a disastrous debate performance in late June and amidst poor polling numbers. Biden quickly endorsed then-Vice President Kamala Harris, who quickly became the Democratic nominee and the first woman of color nominated for President of the United States by a major party. Harris ran a 107-day campaign, and ultimately lost the popular vote to Trump by 1.5 percentage points (48.4% to 49.9%) and the electoral college 226 to 312. I've always been *curious* about what would have happened if Biden had remained the Democratic nominee. Would Trump have won 400 electoral votes, like Pod Save America claimed Biden's internal polling suggested before he exited the race? Or would Biden, who defeated Trump in 2020, have managed to defeat him again?

As shown in Figure 1, Biden was trailing Trump in the polls by around 3 percentage points when he exited the race on July 21, 2024. Trump continued to steadily increase his polling advantage over Biden after the June 27, 2024 debate. For context, former Secretary of State Hillary Clinton lost the electoral college 232 to 306 to Donald Trump in 2016, despite winning the popular vote by just under 2 percentage points. This leads me to my **question of interest: What would Trump's popular vote percentage point margin have been if Biden had remained the Democratic nominee?** This assignment does not *actually* answer that question, but it is an opportunity for me to apply what I learned about GAMs in STA303 and about Time Series Forecasting by reading Duke Professor Robert Nau's [3] Statistical forecasting: notes on regression and time series analysis.

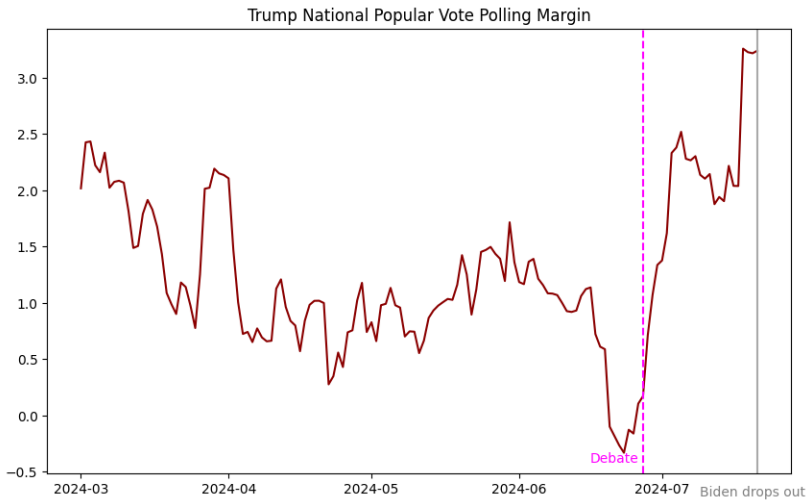

Figure 1: Trump observed percentage-point margin over Biden in 2024 presidential election cycle polls up to Biden's withdrawal from the race on July 21, 2024.

2 Generalized Additive Model (GAM) approach

In STA303, I worked with the `mgcv` package in R to fit Gaussian and non-Gaussian GAMs and Generalized Additive Mixed Models (GAMMs) to various time trends datasets in order to identify trends over time, seasonality, and produce forecasts + prediction intervals. This assignment gave me the chance to apply that knowledge in Python.

The following GAM was fitted using Python's [1] `pygam` package to the observed polling data to capture Trump's margin over Biden:

$$\begin{aligned} \mathrm{Y}_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + f(\text{date}) + \epsilon_i \end{aligned}$$

Where:

- Y_i is Trump's margin over Biden in polls from day i.
- β_0 is the intercept.
- $\epsilon_i \sim N(0, \sigma^2)$ are the residuals.

As shown in Figure 2, my GAM approach estimates that Biden would have lost the popular vote by significantly more than Harris did if he had remained in the race. My GAM estimate is limited because it forecasts 107 days into the future; unsurprisingly, the 95% prediction interval for Trump's popular vote margin on election day is wide. As I learned in STA303, the further into the future a GAM has to forecast, the greater its uncertainty (wider prediction interval) will be. Despite all the uncertainty in this estimate, Trump's actual popular vote margin of victory over Harris (1.5 percentage points) is outside of the prediction interval for all 107 days.

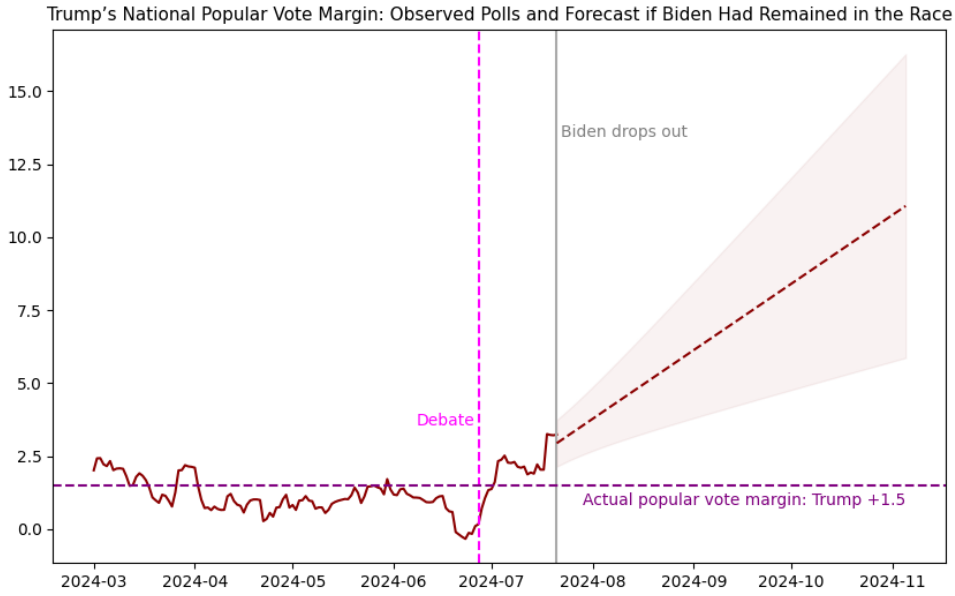

Figure 2: GAM forecast of Trump's popular vote advantage on election if Biden had remained in the race. The solid red lines show Trump's observed margin in 2024 presidential election cycle polls until Biden withdrew on July 21, 2024. The dotted red line shows Trump's estimated margin over the 107 days between when Biden withdrew until election day (November 5, 2024). The light pink shaded area is the 95% prediction interval (there is a 95% probability that Trump's actual margin would have fallen within this range). Based on the GAM estimate, Trump would have won the popular vote on election day by 11.07 percentage points, with a 95% prediction interval of 5.87 to 16.27 percentage points. The prediction interval widens as the time since Biden exited the race increases. The horizontal purple line shows Trump's actual margin; he ended up defeating Harris 49.9% to 48.4% in the popular vote, by a margin of 1.5 percentage points.

3 ARIMA

In this section, I apply what I learned about ARIMA models for time series forecasting to estimate Trump's popular vote win margin if Biden had remained in the race. I use the FiveThirtyEight uncorrected polling data up until July 21, 2024 as my observed data. I used the Python `statsmodels` package to fit an ARIMA(1, 1, 1) model and the `matplotlib` package to display my results. I used the course website titled *Statistical forecasting: notes on regression and time series analysis*, prepared by Robert Nau at Duke University to learn about Time Series analysis and ARIMA models [3].

I began my ARIMA analysis by checking whether my observed polling data is stationary. I performed the ADF test using the Python `statsmodels` package. I obtained the code used for my ADF testing and autocorrelation plots from Gültekin's Medium article titled *Time Series Analysis and Forecasting with ARIMA* [4]. As shown in Figure 3, the ADF test resulted in an ADF statistic of - 1.969 and a p-value of $0.3 > 0.05$, which indicates that my data is not stationary and $d > 0$ should be used to account for non-seasonal differences [4].

```
# check that the data is stationary
adf_test = adfuller(trump_margin_data['trump_margin'])
print('ADF Statistic: %f' % adf_test[0])
print('p-value: %f' % adf_test[1])
```

```
ADF Statistic: -1.969164
p-value: 0.300255
```

Figure 3: ADF test results for the time series data that shows Donald Trump's margin over Joe Biden in 2024 presidential election cycle polls conducted before Biden exited the race on July 21, 2024. The p-value of $0.3 > 0.05$ indicates that the data is not stationary and non-seasonal differencing must be included in the model.

The data was then differenced using $d = 1$ to meet the stationary criterion [3]. As shown in Figure 4, the ADF test on the differenced data reslted in an ADF statistic of -10.4967 and a p-value very close to zero, smaller than 0.05. This means that the time series is now stationary and increasing d to 2 is not necessary [4].

```
[ ] # checking to see if d=1 works
    data_diff1 = trump_margin_data['trump_margin'].diff().dropna()
    adf_test = adfuller(data_diff1)
    print('ADF Statistic: %f' % adf_test[0])
    print('p-value: %f' % adf_test[1])
```

```
ADF Statistic: -10.496606
p-value: 0.000000
```

Figure 4: ADF test results for the differenced data using d=1. The p-value of $\approx 0 < 0.05$ shows that the time series data is now stationary and does not need additional differencing.

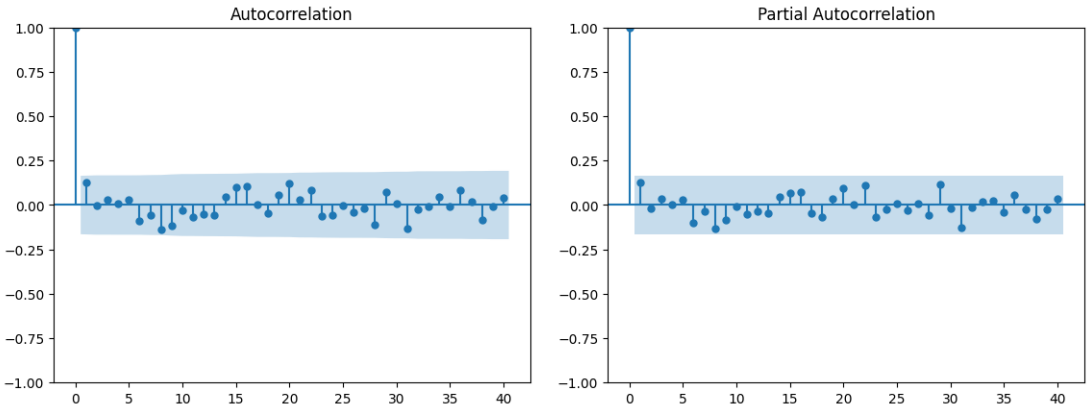

Figure 5: The autocorrelation function (ACF) and partial autocorrelation function (PCF)

The PACF of the differenced series displays a sharp cutoff at the lag-1 and the lag-1 autocorrelation is positive, so an AR term of 1 (p=1) will be used [3, 4]. The ACF of the differenced series has a sharp cutoff at lag 1, so an MA term of 1 (q=1) will also be used [3, 4]. An ARIMA(1, 1, 1) model will be used because it combines the features of autoregressive and moving average models [3] It has the following prediction equation:

$$Y(t) = \mu + Y(t-1) + \phi(Y(t-1) - Y(t-2)) - \theta\epsilon(t-1)$$

An ARIMA(1, 1, 1) model was fit using Python's [1] `statsmodels ARIMA` framework. The results are shown in Figure 6

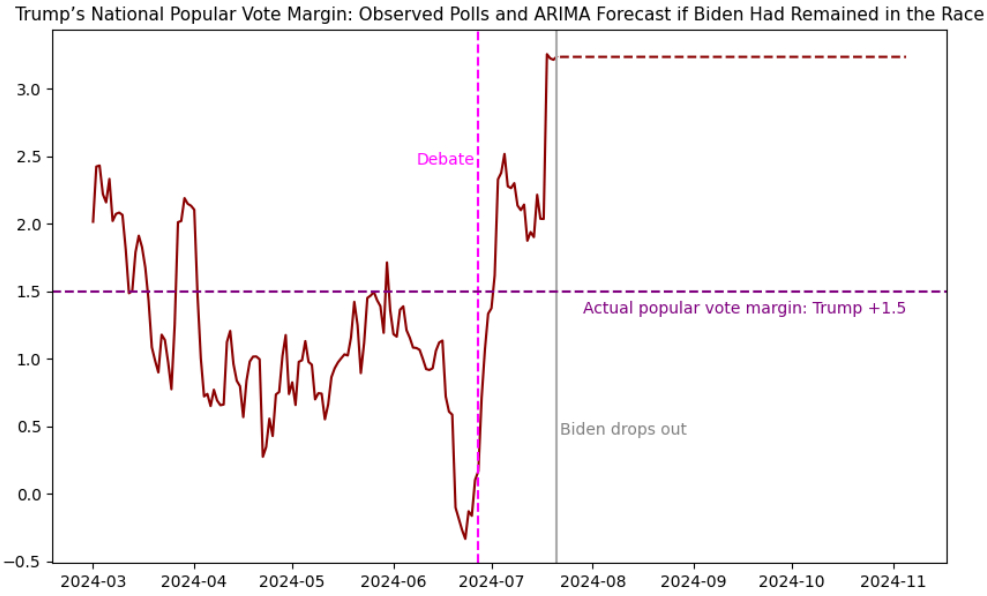

Figure 6: The ARIMA forecast of Trump's national popular vote advantage if Biden had remained in the race estimates that Trump would have won the popular vote by over 3 percentage points. The solid red lines show Trump's observed margin in 2024 presidential election cycle polls until Biden's withdrawal on July 21, 2024. The dotted red line shows the ARIMA(1,1,1) forecast of this margin over the 107 days between Biden's withdrawal and election day.

4 Weaknesses, Limitations, Next Steps

This assignment is solely an academic exercise. It is not a formal estimate of how Biden would have performed in the 2024 election if he had remained the nominee. However, the future, I hope to learn more about classical, Bayesian, and machine learning forecasting methodologies to answer that question. One idea that I have, but still need to learn a lot more about is using recurrent neural networks (RNNs) with long-short-term-memory to forecast polling averages. I learned about both RNNs and LSTM in CSC413 (Neural Networks and Deep Learning), so I hope to expand my knowledge in those areas and how to apply it to election forecasting. As of now, I'm not even sure if or how to get the data for that.

I also faced a couple of data-related limitations when completing this exercise:

- The dataset that I used is not up-to-date. It does not account for Robert F. Kennedy Jr's August 2024 withdrawal and subsequent endorsement of Donald Trump. [2].
- Limited availability of high-quality polling data from the 2024 presidential election cycle.

References

[1] Python Software Foundation, Python, version 3.10, Python Software Foundation, 2021. [Online]. Available: https://www.python.org

[2] FiveThirtyEight, "2024 presidential election poll averages," GitHub repository, 2024. [Online]. Available: https://github.com/fivethirtyeight/data/tree/master/polls/2024-averages

[3] R. Nau, "Statistical forecasting: notes on regression and time series analysis," Fuqua School of Business, Duke University. [Online]. Available: https://people.duke.edu/ rnau/411home.htm

[4] H. Gültekin, "Time Series Analysis and Forecasting with ARIMA," Medium, Mar. 24, 2024. [Online]. Available: https://medium.com/@hazallgultekin/time-series-analysis-and-forecasting-with-arima-8be02ba2665a